A Model for Saliency-Based Visual Attention for Rapid Scene Analysis

Laurent Itti, Christof Koch and Ernst Niebur

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,
VOL. 20, NO. 11, NOVEMBER 1998

Jyothsna Shashikumar Sastry

Overview

- Motivation
  - Introduction to Saliency based Model
- Architecture
  - Extraction of Early Visual Features
  - Saliency Maps
- Comparison with Spatial Frequency Content Models
- Results
- Strengths and limitations
- Summary

Activity

Analyse this scene

MOTIVATION

- What did you see? List the salient objects you identified.
- We can interpret complex scenes in real time
- Focus of Attention (FOA) processes only a subset of the sensory information to reduce complexity of scene analysis

GOAL : Build a model that mimics primate visual attention for static scenes

- Models of Attention
  - Dynamic routing Models
  - Saliency based Models

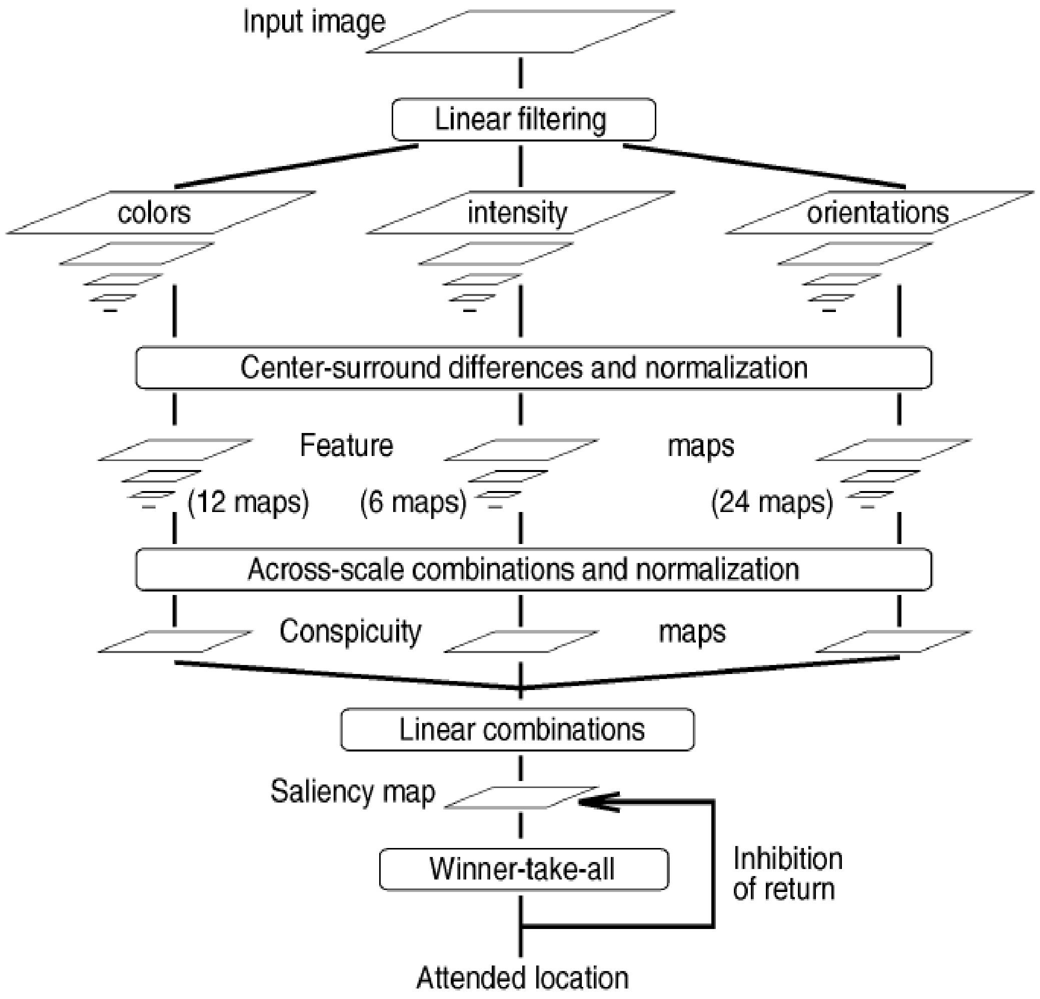
- Based on feature integration theory
- Visual input is decomposed into a set of feature maps
- Different spatial locations within a map compete for saliency
- All feature maps feed into a master “saliency map”

Massively parallel method for the rapid selection of a small number of interesting image locations to be analysed by more complex and time-consuming object-recognition processes

ARCHITECTURE

- Input: Static color images digitized at 640 X 480
- Multiscale feature extraction: Nine spatial scales using Gaussian pyramids ($\sigma \in [0..8]$ where σ is the scale)
- **Center surround** implemented to detect locations that stand out from their surround
    - **Center**: a pixel at scale $c \in \{2, 3, 4\}$
    - **Surround** is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{3, 4\}$.

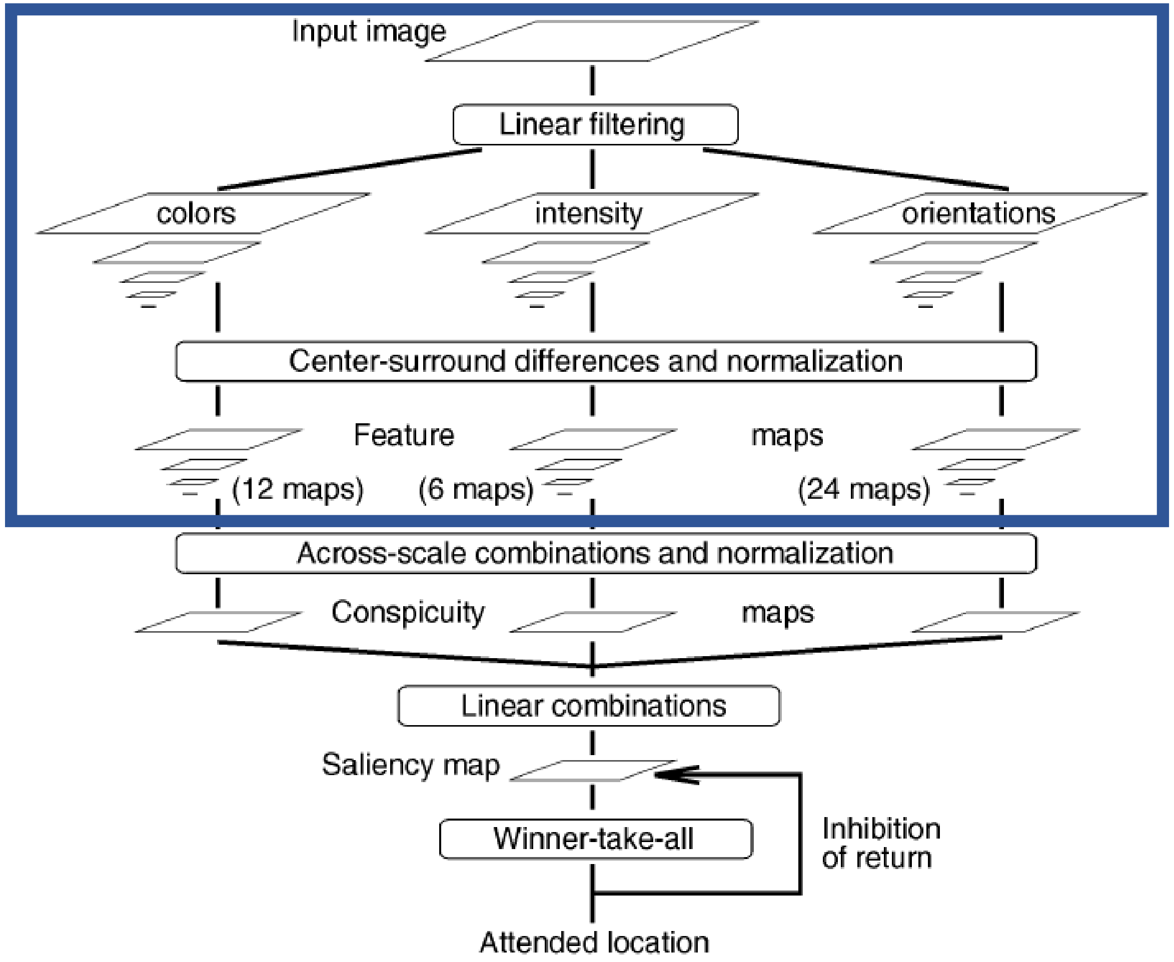
The across-scale difference between two maps, denoted "$\ominus$"



Extraction of Early Visual Features

Linear Filtering and **Center-surround differences and Normalization**

- Intensity
- Color
- Orientation



Extraction of Early Visual Features

Intensity contrast maps

- Mammals are equipped with neurons sensitive to light centers with dark surrounds and dark centers with light surrounds
- Intensity image $I = (r + g + b) / 3$
- Gaussian pyramid $I(\sigma)$ where $\sigma \in [0..8]$
- Normalize each channel by I only where $I > (1/10)$th of I_{max} (maximum I over entire image)
- Intensity contrast maps $\boldsymbol{I}(c, s) = |\boldsymbol{I}(c) \ominus \boldsymbol{I}(s)|$

  $c \in \{2, 3, 4\}$ and $s = c + \delta$, with $\delta \in \{3, 4\}$.

Extraction of Early Visual Features

Color feature maps

- "Color double-opponent" system in the human cortex
- Four broadly-tuned color channels are created
  - $R = r - (g + b)/2,$
  - $G = g - (r + b)/2$
  - $B = b - (r + g)/2$
  - $Y = (r + g)/2 - |r - g|/2 - b$
- Four Gaussian Pyramids from these channels
  - R(σ), G(σ), B(σ) and Y(σ)
- According to the color opponency,

$\boldsymbol{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$

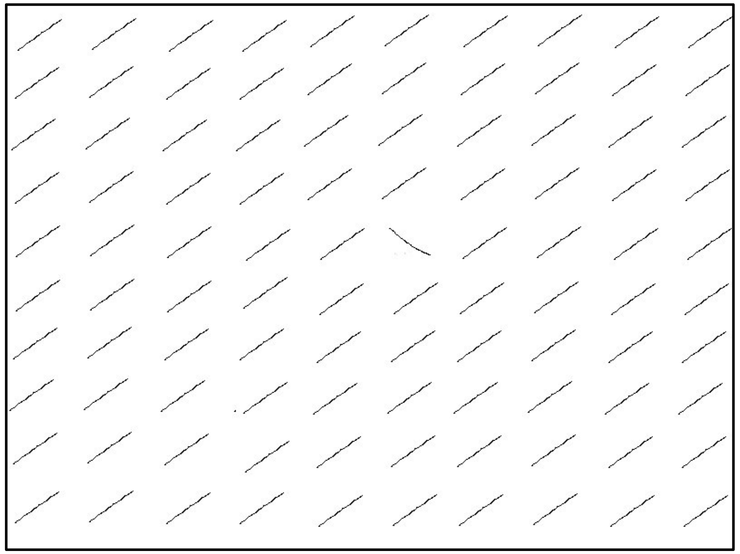
$\boldsymbol{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$

Extraction of Early Visual Features

Orientation maps

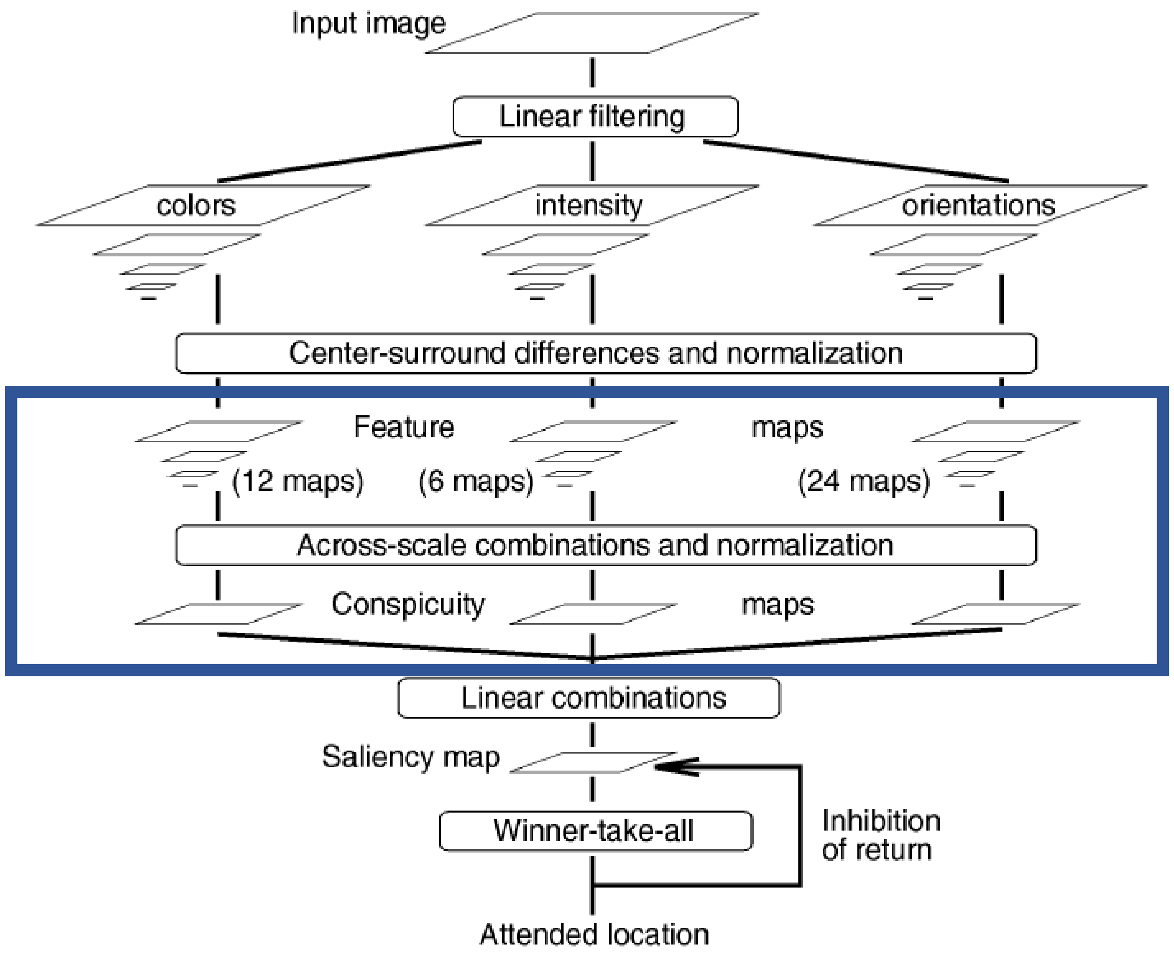
- Human visual attention system has the ability to identify breaks in patterns with the aid of orientation sensitive neurons of the cortex
- Intensity image **I** convolved with an Orientational filter (Gabor filters)
- $\Theta \in \{0^0, 45^0, 90^0, 135^0\}$
- Gabor pyramids at nine scales [0...8]

$\boldsymbol{O}(c, s, \Theta) = |O(c, \Theta) \ominus O(s, \Theta)|$



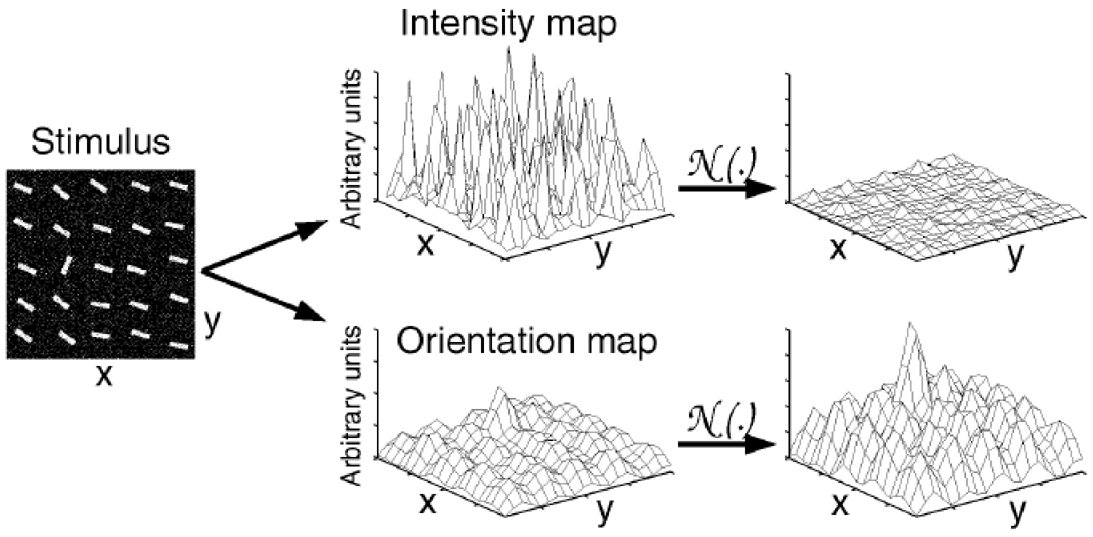
Saliency Maps

- 42 Feature maps:
  - 6 for Intensity
  - 12 for color
  - 24 for orientation
- Difficulty in combining different feature maps?



Saliency Maps

- Normalization operator ***N***(.)
  - Normalize map to a fixed range [0…M]
  - Find location of global maximum M
  - Compute average $\overline{m}$ of the local maxima
  - Multiply map by $(M-\overline{m})^2$



Saliency Maps

- Feature maps are combined into three "conspicuity maps" by across-scale additions at σ = 4 of the saliency map

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}\left(I(c,s)\right)$$

$$\bar{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[\mathcal{N}\left(\mathcal{RG}(c,s)\right) + \mathcal{N}\left(\mathcal{BY}(c,s)\right)\right]$$

$$\bar{O} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}\left(O(c,s,\theta)\right)\right)$$

- The three conspicuity maps are averaged into the final input **S** to the saliency map

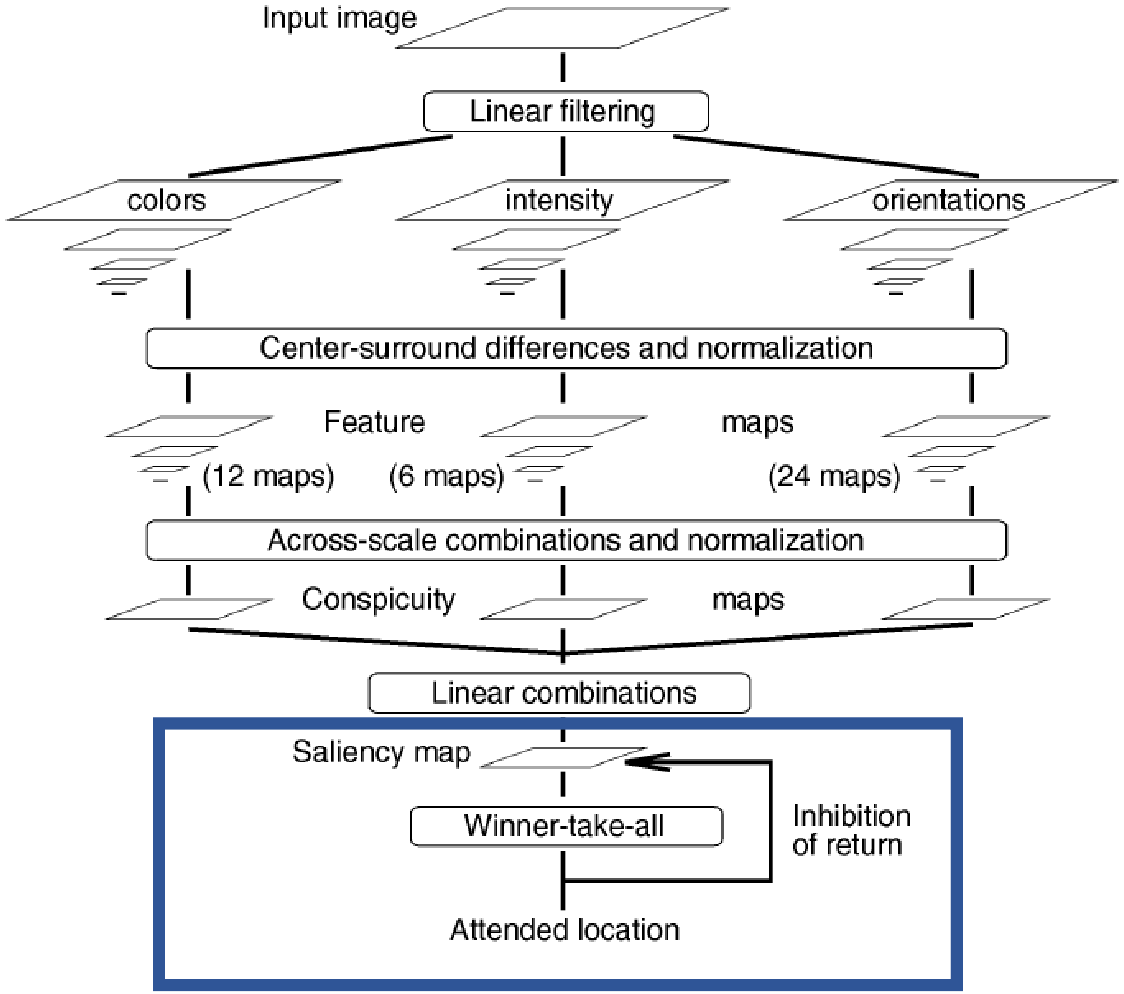
$$\mathbf{S} = \left(\mathbf{N}(\bar{I}) + \mathbf{N}(\bar{C}) + \mathbf{N}(\bar{O}) \right) / 3$$

Saliency Maps

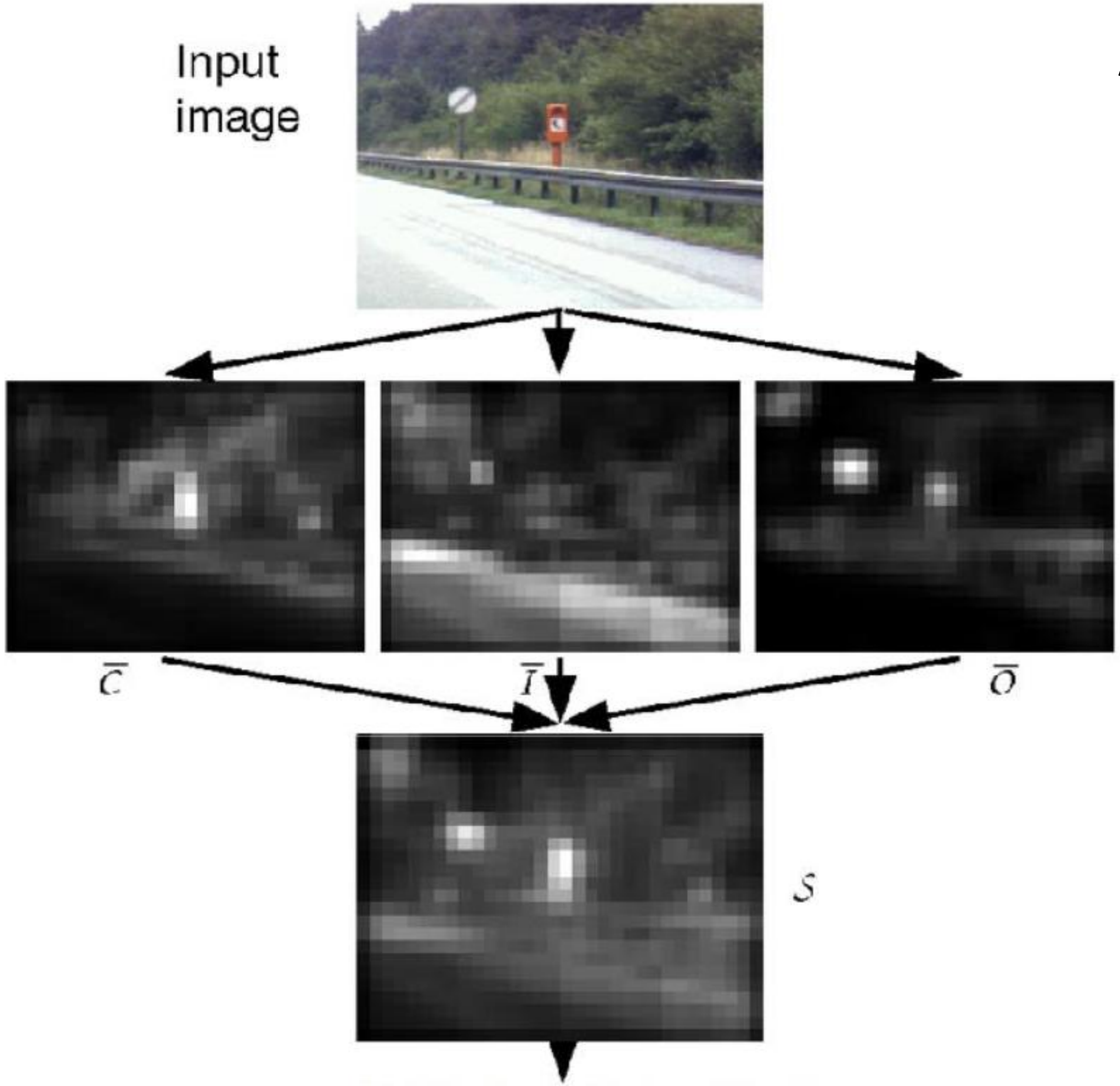
Model the Saliency Map as a 2D layer of *integrate and fire* neurons at scale 4

Working:

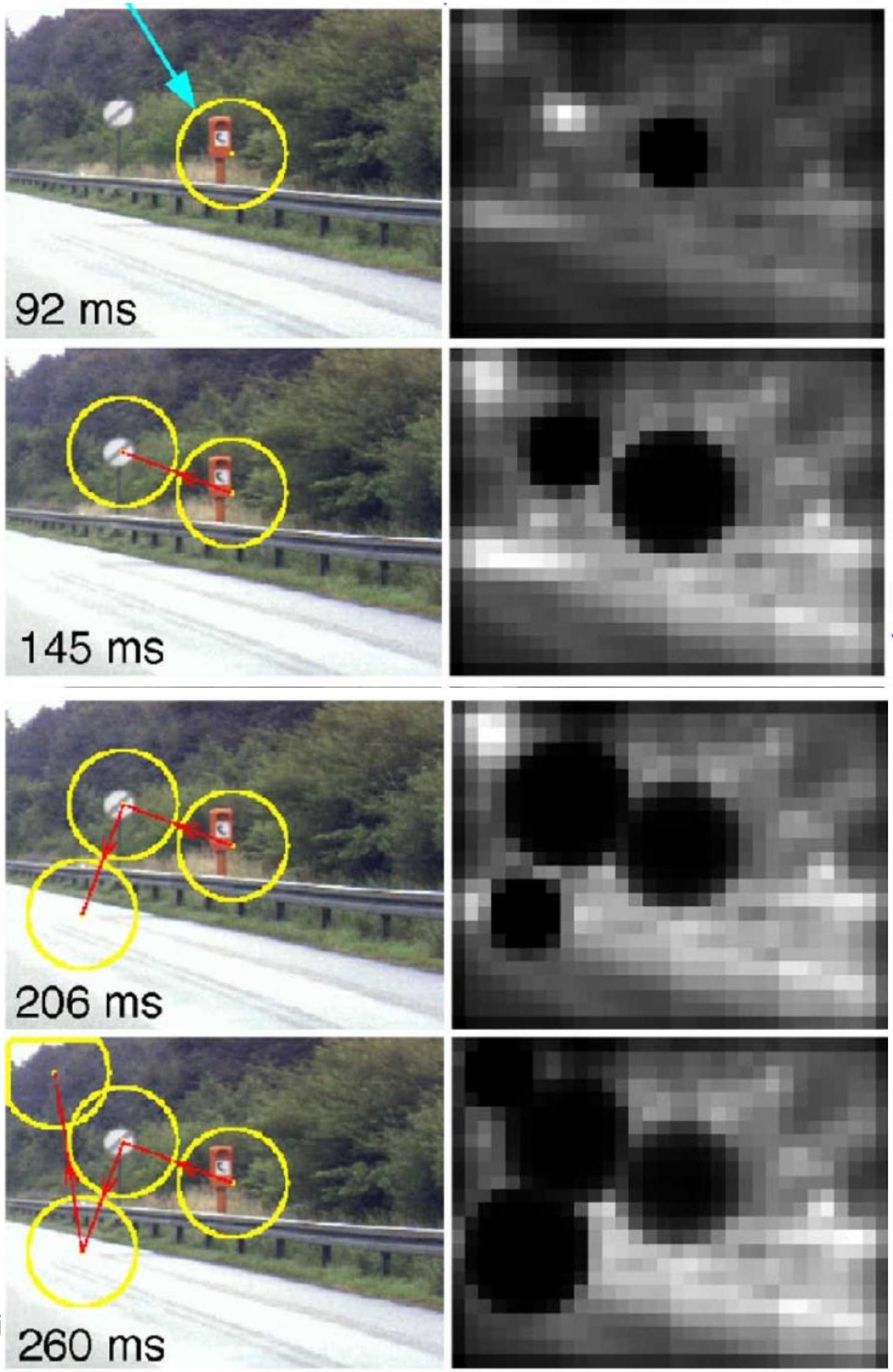
- Neurons in the SM receive excitatory inputs from *S*
- Each SM neuron excites its corresponding Winner Take All neuron.
- The first "winner" to reach a threshold fires.
- FOA shifted to this location
- All WTA neurons are reset
- SM neurons are reset in the location of FOA – aiding next salient location to become the winner



Input image



Focus of Attention



Saliency maps

Comparison with Spatial Frequency Content Models

- At a given image location, a 16 X 16 image patch is extracted from each I(2), R(2), G(2), B(2), and Y(2) map
- 2D Fast Fourier Transforms (FFTs) are applied to the patches
- The SFC measure is the average of the numbers of nonnegligible coefficients in the five corresponding patches

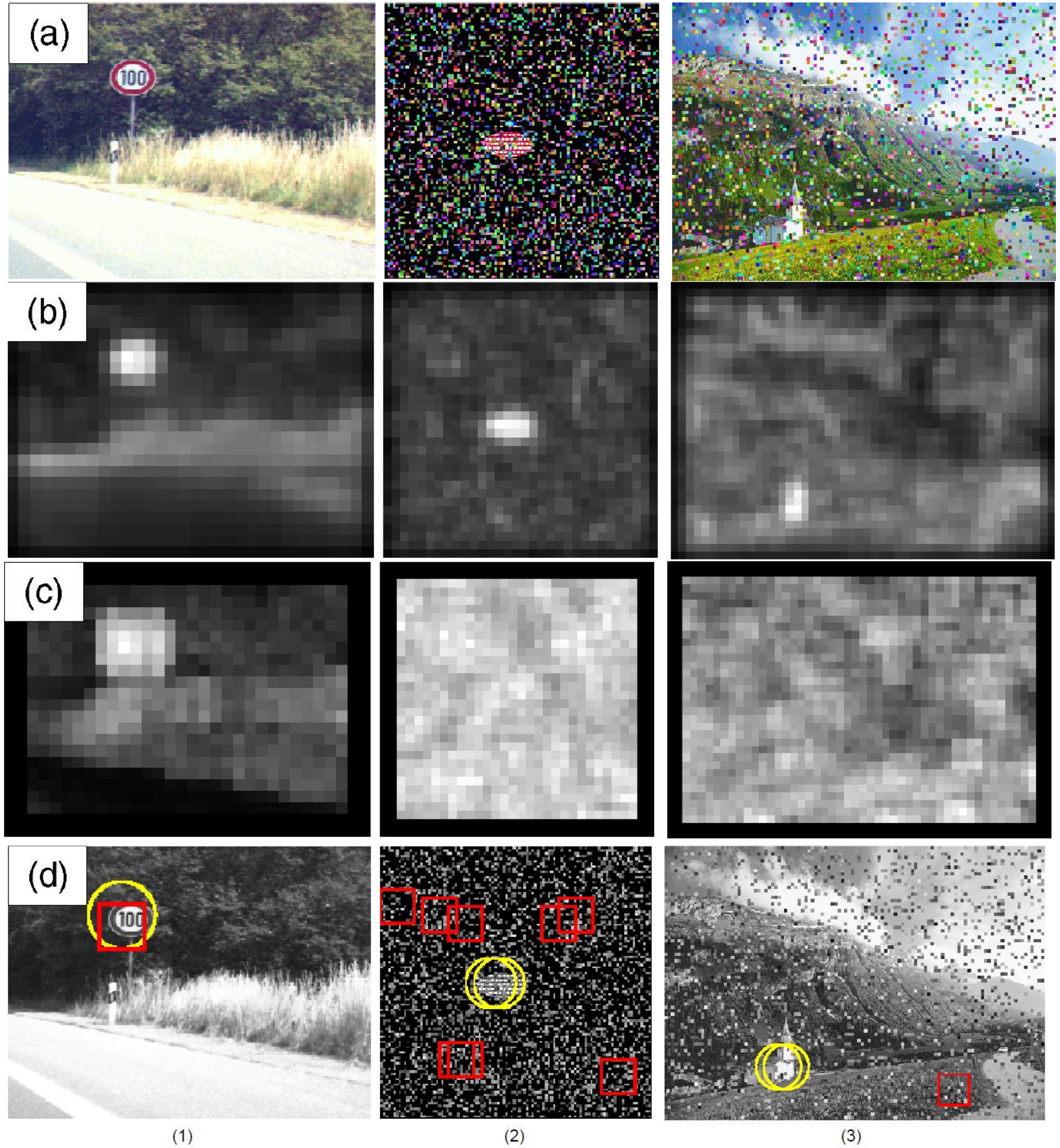
- Results show that our model is superior to spatial frequency content (SFC) based models in the presence of noise

Input Images

Saliency maps

SFC maps

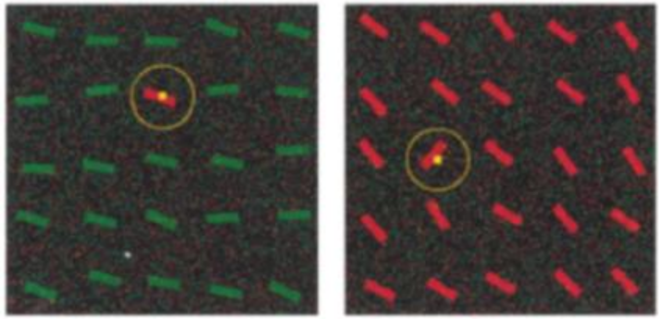
Output salient
Locations
SM(Yellow circle)
SFC(Red square)



RESULTS

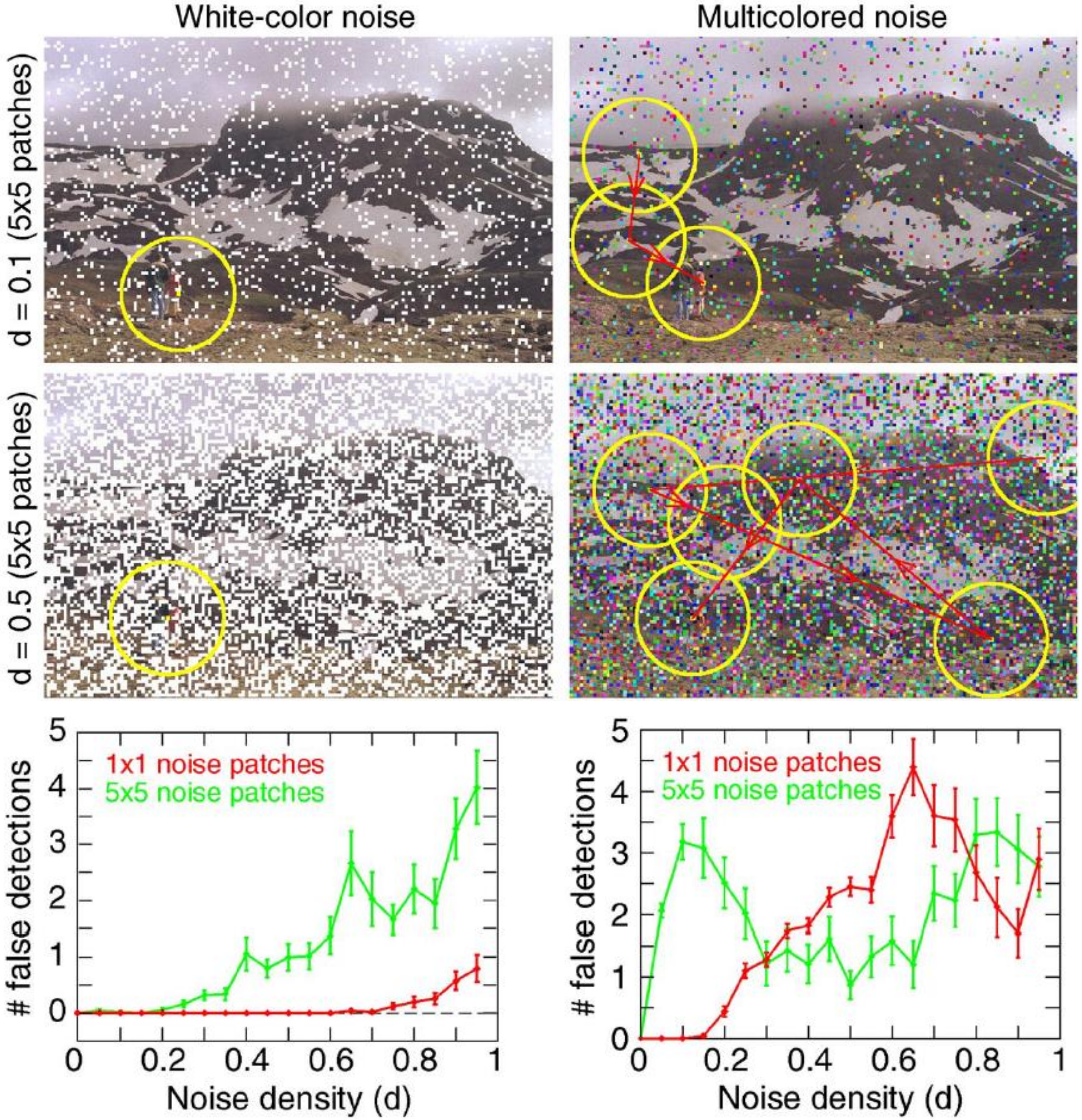
General Performance

- Extensively tested with artificial images to ensure proper functioning
- Robust to addition of noise
- Reproduces human performance for pop-out tasks



Color pop-out Orientation pop-out

- Tested with real images: attended locations were the objects of interest


White-color noise
Multicolored noise
d = 0.1 (5x5 patches)
d = 0.5 (5x5 patches)
1x1 noise patches
5x5 noise patches
false detections
Noise density (d)
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5

Strengths

✓Mimics the properties of primate vision

✓Despite its simple architecture and feed-forward feature-extraction mechanisms, the model is capable of strong performance with complex natural scenes

✓Massively parallel implementation for the rapid selection of a small number of interesting image locations

✓Allows real time operation

Limitations

- Only object features explicitly represented in at least one of the feature maps can lead to pop-out
- Fails to detect targets salient for unimplemented feature types (e.g., T junctions or line terminators)
- Cannot reproduce contour completion or closure
- No magnocellular motion channel

SUMMARY

- A conceptually simple computational model for saliency-driven visual attention.
- Architecture is inspired by biological insights
- Efficient in reproducing some of the performances of primate visual systems.
- The efficiency of this approach for target detection critically depends on the feature types implemented.
- Model can be easily tailored to arbitrary tasks using dedicated feature maps.

REFERENCES

- **A Model of Saliency-Based Visual Attention for Rapid Scene Analysis** Laurent Itti, Christof Koch, and Ernst Niebur
- https://github.com/mbanani/attend
- Image source: Google

Thank you!


Q
A